Designing a Data Analysis Subsystem for Predicting the Properties of Antifungal Antibiotics

Eldar E. Musayev, Tamara Chistyakova, Vera A. Kolodyaznaya, and Valery V. Belakhov

Abstract In this chapter, we describe the algorithms for data processing applied as part of an intellectual analysis subsystem of a software system for predicting and researching the properties of antifungal antibiotics. These include models for predicting toxicity based on assays as well as acute oral toxicity. The mathematical models were trained, tested, and validated on different sets of antifungal antibiotic data. Testing showed the models' accuracy and viability for predicting antifungal antibiotics' properties.

Keywords Software development · Antifungal antibiotics · Mathematical models · Machine learning

1 Introduction

Fungal infections are one of the most important issues in healthcare. The number of fungal infections is growing as a result of, among other reasons, continuing environmental pollution, an increase in background radioactivity, improper application of broad-spectrum antibiotics, growing use of cytostatic and immunosuppressive drugs, and the appearance of more and more frequent antifungal drug resistance [1–3]. Among these infections, invasive mycoses are becoming a more and more important medical concern due to the growing number of immunocompromised patients [4–6]. The number of currently available and approved systemic antifungals is insufficient [7–9], and the progress of developing novel antifungal drugs is not fully proportional

E. E. Musayev (✉) · T. Chistyakova
Department of Computer-Aided Design and Control, Saint Petersburg State Institute of Technology (Technical University), Moskovsky Ave, 26, Saint Petersburg 190013, Russia

V. A. Kolodyaznaya
Department of Biotechnology, Saint Petersburg State Chemical-Pharmaceutical University, Professor Popov Str., 14, Saint Petersburg 197376, Russia

V. V. Belakhov
Department of Chemistry, Technion - Israel Institute of Technology, 3200008 Haifa, Israel
e-mail: chvalery@technion.ac.il

A. G. Kravets et al. (eds.), *Society 5.0: Human-Centered Society Challenges and Solutions*, Studies in Systems, Decision and Control 416,
https://doi.org/10.1007/978-3-030-95112-2_5

to the rate of growth of antifungal diseases, which include invasive fungal infections that are an existential and growing problem for modern healthcare [10–12]. Effective use of antifungal drugs to treat various mycoses is an important factor in the fight against antifungal infections.

One of the main issues affecting drug research is the cost of research and development, which can reach as high as 2.5 billion dollars [13]. The time it takes to develop a new drug is also a key issue, as a great deal of time is lost on drugs that ultimately do not pass pre-clinical or clinical trials.

One of the modern approaches to developing novel highly-effective low-toxicity antifungal drugs with improved medical, biological, and biopharmaceutical properties is the chemical modification of existing antifungal drugs, chief among those polyene macrolide antibiotics [14–16]. In this chapter, we discuss a specific class of antibiotics: polyene macrolide antibiotics (PMA), which make up approximately a quarter of all existing antifungal antibiotics. The chemical structure of a PMA consists of a macrolide ring that contains conjugated double bonds on one side (forming the lipophilic side of the molecule) and a number of hydroxyl and keto groups on the other (forming the hydrophilic side of the molecule). Their biological target is ergosterol, one of the components of a pathogenic fungi phospholipid membrane.

Amphotericin B is the drug of choice (gold standard) among all known PMA due to its high antifungal activity against the vast majority of known clinical forms of mycoses. PMA derivatives (PMAD) are chemically modified versions of existing PMA drugs which retain the biological activity of the initial drug while having lower toxicity. They can be an important topic of research in the fight against fungal drug resistance [17–19].

Software engineers can support the process of PMAD research by creating a software system that can predict antifungal activity and toxicity on the basis of the chemical structure of a molecule.

The goal is the development of models and a software solution providing those models that can reduce antifungal drug research time and cost by selecting such PMAD that have lower toxicity while retaining their ability to bind to the biological target. Using such a program, a researcher can check the toxicity and antifungal activity of a potential PMAD, and select such PMAD to go to pre-clinical trials that have more favorable traits. The program helps to cut time and other resource expenditure for pre-clinical and clinical trials of PMAD that lack the desired pharmaceutical properties.

2 Description of the Software System

The software system contains interfaces for researchers, experts, and database administrators. It includes an intelligent data analysis subsystem, a subsystem of synthesis step selection, and databases providing them with the data they require to function (see Fig. 1).

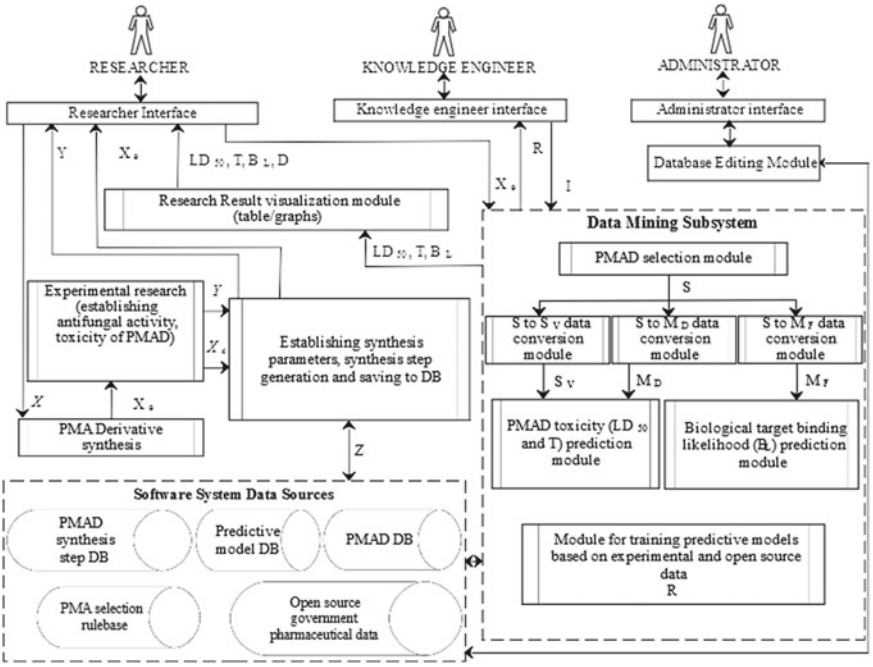

Fig. 1 Architecture of the software system for predicting and researching antifungal antibiotics' properties

Where LD50—the lethal dose for half of the population (mg drug/kg, oral intake, rats), T—a vector of predicted results of assays corresponding to toxicity signaling reactions, BL—the likelihood of binding to the biological target (%), D—the graph representation of the molecule, I—additional data to train the neural networks on, R—the results of this training process (AUC, MSE), S—the SMILES notation representation of the molecule's structure, SV—a vectorized version of that notation, MD—a vector of molecular descriptor values generated from the SMILES notation representation, MF—Morgan's molecular fingerprint bit vector for the molecule, X—description of the initial PMA, Xe—the result of modifying the structure of that PMA to create a PMAD, Y—the experimentally derived values acquired by testing that PMAD, and Z—the PMAD synthesis steps.

The data analysis subsystem consists of one acute toxicity model based on gradient boosted decision trees, 12 recurrent neural networks modeling one property each based on embedded vector representations of the elements of the SMILES notation of the molecule, and a deterministic algorithm for predicting biological activity based on pharmacophore filtering.

3 Data Analysis Subsystem Components

The acute toxicity model utilizes the gradient boosted decision tree model catboost in order to predict toxicity. It is trained on data retrieved from ChemIdPlus [20] in the form of tsv data (approximately 6000 values). Of note is that, prior to predicting the value, we multiply it by a normalized (0, 1] value of its logP in order to adjust somewhat for absorption differences due to lipophilicity. The data is input as a SMILES string, then processed using RDKit [21], which also provides us with the descriptors we use and the RDKit molecular fingerprint that also serves as input. The data is then fed into a gradient-boosted decision tree (catboost) model. The model's hyperparameters are as follows: iterations: 50,000, depth: 6, od_type: 'Iter', od_wait: 500, learning_rate: 0.07, random_strength: 40, l2_leaf_reg: 100, rsm: 0.3.

The predicted values are divided by the normalized logP value. The normalizer model is stored alongside the catboost model.

The assay-based toxicity prediction has a pre-processing step. First, we use the Chembl database [22] to attain approximately 1.7 million SMILES representations of valid molecules. We then determine all of the unique elements the SMILES notation consists of and one-hot encode them. We then utilize a skip-gram variant of word embeddings on these elements. The window size is 11 (5 to each side of the predicted element) and the embedded vector has 15 elements. The skip-gram variant of neural network encoding for embedded vector attainment is presented in Fig. 2.

We limit predictions to SMILES notations of at most 300 elements. If a SMILES notation is shorter than 300 elements, we append zero vectors to ensure all inputs are of identical (300, 15) shape. The utilized neural network consists of a bi-directional GRU layer, represented in Fig. 3. The network is trained on the tox21 dataset [23].

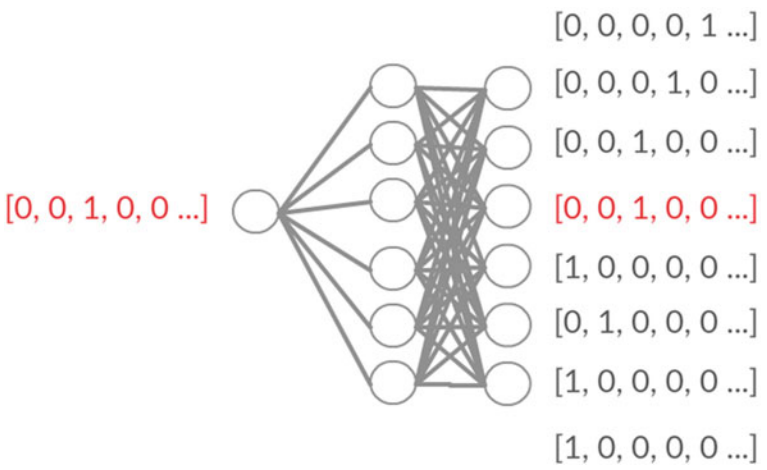

Fig. 2 Skip-gram encoding to attain an information-rich embedded vector (represented here as the hidden layer)

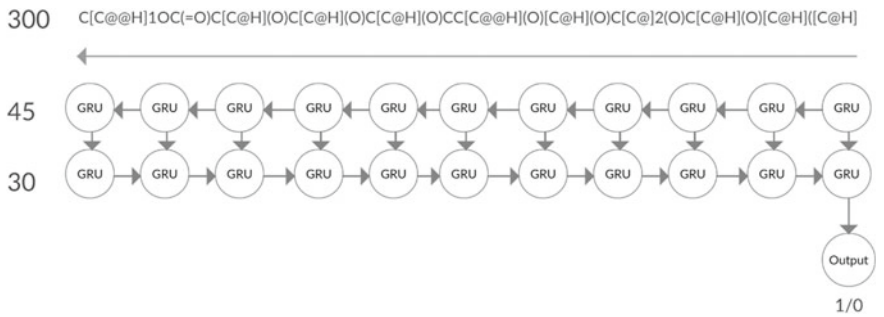

Fig. 3 Representation of a bi-directional GRU model for predicting one assay-based toxicity indicator. The figure represents the SMILES notation in its character form, though in the actual input each element is replaced by its embedded vector

The biological activity pharmacophore filtering model is a deterministic algorithm. First, RDKit fingerprints of a subset of polyene macrolide antifungal antibiotics are generated. Then, these are combined into a single fingerprint in such a way that only those features that exist in each of them are left in the resulting pharmacophore. This new fingerprint is treated as the minimum set of features required for antifungal activity. The algorithm to predict the probability of antifungal activity is as follows:

$$p = (\Sigma 2048_i = 1\ M_i * PH_i)/(\Sigma 2048_i = 1\ PH_i),$$

where p—the predicted value, [0, 1], corresponding to the probability that the input molecule will show antifungal activity;

Mi—the i-th value of the 2048-bit vector, corresponding to the presence or absence of a structural element of the researched molecule;

PH_i—the i-th value in the 2048-bit vector, corresponding to the presence or absence of a structural element of the pharmacophore;

In order to use this algorithm to classify researched molecules as having or lacking antifungal activity, a cutoff value is used. The selected cutoff value is 0.95, meaning that 95% of all structural elements of the pharmacophore must be present in any researched molecule for it to be marked as an antifungal antibiotic.

4 Interpretation and Discussion of Research Results

The acute toxicity model's root means squared error was 58 mg/kg (LD50, oral intake, rats). This was acceptable given that the molar weight of antifungal antibiotics tends to be 600+ g/mol. The assay-based toxicity prediction utilizing the tox21 dataset results is presented in Table 1.

Table 1 Tox21 modeling results

Assay	AUC	f-1
NR-AR	0.783569	0.969507
NR-AR-LBD	0.84447	0.981438
NR-AhR	0.88377	0.938144
NR-aromatase	0.799273	0.954627
NR-ER	0.715153	0.875431
NR-ER-LBD	0.826407	0.953131
NR-PPAR-gamma	0.767862	0.96973
SR-ARE	0.864636	0.888909
SR-ATAD5	0.859089	0.974173
SR-HSE	0.870134	0.969882
SR-MMP	0.912109	0.91518
SR-p53	0.854033	0.967989

The biological activity prediction pharmacophore filter was tested on a set of antifungal antibiotics as well as a set of drugs that are not antifungal antibiotics. With the selected cutoff point of 0.95, all of the antifungal antibiotics were correctly classified as such, and none of the non-antifungal drugs were classified as antifungal drugs.

5 Conclusion

We proposed an approach to designing a data analysis subsystem for a software system for predicting and researching the properties of antifungal antibiotics. These include gradient-boosted decision tree models, recurrent neural networks, and non-statistical algorithms. The software solution is configurable to various types of antifungal antibiotics, and its models can be trained on more antifungal antibiotic derivatives data to improve their accuracy. Testing was performed using sets of existing antifungal antibiotics as well as a number of recently synthesized novel antibiotics [14–16, 18, 19]. Testing supports the applicability of the system for predicting antifungal antibiotics' properties.

References

1. Jucker, E.: Antifungal Agents: Advances and Problems, Special Topic: Progress in Drug Research. Birkhaeuser Verlag, Basel (2003)
2. Coste, A.T., Vandeputte, P.: Antifungals: From Genomics to Resistance and the Development of Novel Agents. Caister Academic Press, Norfolk (2015)

3. San-Blas, G., Calderone, R.A.: Pathogenic Fungi: Insights in Molecular Biology. Caister Academic Press, Norfolk (2008)
4. Sergeev A.U., Sergeev U.V.: Candidiasis. Nature of infections, mechanisms of aggression and defense, laboratory diagnostics, clinics and treatment. Triada-X, Moscow (2001)
5. Sergeev A.U., Sergeev U.V.: Fungal Infections. Manual for doctors. BINOM, Moscow (2008)
6. Kozlov, S.N., Strachunskiy, L.S.: Modern antimicrobial chemotherapy. OOO "Medicinskoye infromacionnoye agenstvo", Moscow (2009)
7. Reiss, E., Shadomy, H.J., Lyon, G.M.: Fundamental Medical Mycology. Willey-Blackwell, Hoboken (2011)
8. Sillivan, D.J., Morgan, G.P.: Human Pathogenic Fungi: Molecular Biology and Pathogenic Mechanisms. Caister Academic Press, Norfolk (2014)
9. Omura, S.: Macrolide Antibiotics: Chemistry, Biology and Practice. Academic Press, New York (2002)
10. d'Enfert, C., Hube, B.: Candida: Comparative and Functional Genomics. Caister Academic Press, Norfolk (2007)
11. Masayuki, M., Gomi, K.: Aspergillus: Molecular Biology and Genomics. Caister Academic Press, Norfolk (2010)
12. Veselov, A.V., Kozlov, R.S.: Invasive candidiasis: current aspects of epidemiology, diagnosis, therapy and prevention in different categories of patients. Clin. Microbiol. Antimicrob. Chemother. (18), 1–104 (2016)
13. Tufts Study Finds Big Rise In Cost of Drug Development, https://cen.acs.org/articles/92/web/2014/11/Tufts-Study-Finds-Big-Rise.html (2017). Accessed 10 May 2017
14. Solovieva, S.E., Olsufyeva, E.N., Preobrazhenskaya, M.N.: Chemical modification of antifungal polyene macrolide antibiotic. Russ. Chem. Rev. **80**(2), 115–138 (2011)
15. Omelchuk, O.A., Tevyashova, A.N., Shchekotikhin, A.E.: Recent advances in antifungal drug discovery based on polyene macrolide antibiotics. Russ. Chem. Rev. **87**(12), 1206–1225 (2018)
16. Belakhov, V.V., Garabadzhiu, A.V., Chistyakova, T.B.: Polyene macrolide antibiotic derivatives: preparation, overcoming drug resistance, and prospects for use in medical practice. Pharm. Chem. J. **52**(11), 890–901 (2019)
17. Shavit, M., Pokrovskaya, V., Belakhov, V., Baasov, T.: Covalently linked kanamycin–Ciprofloxacin hybrid antibiotics as a tool to fight bacterial resistance. Bioorg. Med. Chem. **25**(11), 2917–2925 (2017)
18. Belakhov, V.V., Garabadzhiu, A.V., Kolodyaznaya, V.A.: Search for new derivatives of polyene macrolide antibiotics as potential antifungal agents for the delaying of drug resistance and treatment of invasive mycoses. Izvestiya Sankt-Peterburgskogo gosudarstvennogo tehnologicheskogo instituta (tehnicheskogo universiteta) (30), 31–41 (2015)
19. Belakhov, V.V., Garabadzhiu, A.V., Chistyakova, T.B.: Hydrophosphoryl derivatives of tetramycin B: Design, synthesis, biological activity and development of intellectual computer system. Phosphorus Sulfur Silicon Relat. Elem. **194**(4–6), 442–443 (2019)
20. Tomasulo, P.: ChemIDplus-super source for chemical and drug information. Med. Ref. Serv. Q. **21**(1), 53–59 (2002)
21. Landrum, G.: RDKit documentation. Release **1**, 1–79 (2013)
22. Gaulton, A., Bellis, L.J., Bento, A.P., Chambers, J., Davies, M., Hersey, A., Light, Y., McGlinchey, S., Michalovich, D., Bissan, A., Overington, J.P.: ChEMBL: a large-scale bioactivity database for drug discovery. Nucleic Acids Res. **40**, 1100–1107 (2012)
23. Attene-Ramos, M.S., Miller, N., Huang, R., Michael, S., Itkin, M., Kavlock, R.J., Austin, C.P., Shinn, P., Simeonov, A., Tice, R.R., Xia, M.: The Tox21 robotic platform for the assessment of environmental chemicals–from vision to reality. Drug Discov. Today **18**(15–16), 716–723 (2013)